International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon-20..)

Demystifying IoT Network Intrusion Detection : Assessing ML Algorithms with the Aid of Explainable AI

Tasfia Zaima[1], Tabassum Ibnat Ena[1], Md. Tamim Ikbal[1] and Abu Sayed Md. Mostafizur Rahaman[2]

[1] *Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh*
[2] *Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh*

E-mail address: tasfia.zaima000@gmail.com, tabassumibnat245@gmail.com, tamimmd12348@gmail.com, asmmr@juniv.edu

Abstract: Intrusion Detection Systems (IDSs) are pivotal for network security; while machine learning based IDSs surpass traditional models in effectiveness, their growing complexity poses transparency challenges. This study uses the UNSW-NB15 dataset to train the ML algorithms, aiming to demystify the complexity of IoT network intrusion detection. The Explainable Artificial Intelligence (XAI) framework is used to improve model comprehensibility and transparency. Scikit-Learn, ELI5 Permutation Importance, and Local Interpretable Model-Agnostic Explanation (LIME) are applied to analyze the performance of many ML algorithms. This study also investigates the influence of dataset balancing on the performance metrics of various ML algorithms. SVM accuracy rose from 86 to 88 percent, while Random Forest and CatBoost accuracy climbed from 90 to 92 percent after balancing. Ensemble combinations also showed improved performance. ELI5 and LIME were then applied to the ML algorithms. The methodology presented in this paper offers a valuable toolkit for cybersecurity experts, empowering them to make informed decisions in the face of evolving cyber threats. The findings support the integration of XAI approaches with conventional ML systems to improve interpretability in cybersecurity applications. This study enhances IDSs for IoT networks by bridging the gap between ML-based prediction performance and the need for transparent and interpretable decision-making.
Keywords: Explainable AI (XAI), Random Forest (RF), SVM, CatBoost, ELI5, LIME, Permutation Importance (PI)

1. INTRODUCTION

IoT networks, vast reservoirs of user data, are becoming more and more integrated in our daily lives and are critical to the delivery of many crucial services. Our daily activities rely so largely on the internet that it has become a target for hackers. Intrusion detection systems (IDSs) are hardware or software solutions that automate the act of keeping an eye on events happening within a network or computer system and interpreting them to look for indicators of potential security issues. In the last several years, the frequency and intensity of network attacks have increased, making intrusion detection systems an essential component of every organization's security infrastructure. They accomplish this by employing a range of machine learning algorithms that, if any suspicious activity or possible cyberattack is detected, produce notifications (1).Some of the basic drawbacks of the conventional IDS are illustrated by the system, including low power consumption when in use, an absence of well-established IoT protocols and architecture, and a limited computing capability. These systems are good at differentiating between attack and normal behaviors and in identifying apprehensive behaviors and activities. However, their “black box” design makes it strenuous to completely comprehend the decision-making processes that underlie the predictions and attack classifications. This deficiency is essential for creating powerful information assurance plans and optimal cybersecurity assessments. One of the effective and potential ways to identify this issue is to incorporate Explainable AI (XAI) (2) which makes it easier for the analysts to comprehend the reasoning behind the algorithm’s decision. This study portrays an extensive overview of Machine Learning (ML) algorithms, with a particular emphasis on Random Forest (RF), Support Vector Machines (SVM), and CatBoost, and assesses how well they work in the context of IoT network security. The research uses XAI methodologies by using various combinations of ensemble techniques for these classifiers and by utilizing Python tools that enhances the explainability and transparency. Explain Like I’m 5 (ELI5) and Local Interpretable Model-Agnostic Explanations (LIME) are used to visualize feature importance, boosting the models’ explainability in classification prediction scenarios. LIME and ELI5

include feature ranking that makes it easier to assess the in-depth effectiveness and efficiency of the selected machine learning algorithms. The problem at hand involves conducting a comprehensive evaluation of the real-world applicability of an IDS enhanced with XAI (3). The specific objectives of this research are-

- To evaluate the real-world applicability of the XAI enhanced IDS.
- To apply XAI to modify the ML algorithms to increase transparency and provide an explanation for the algorithmic choices.
- Using XAI to rank the features according to their importance.
- To assess ML algorithm performance with the aid of XAI.

In this journal, focus is given on the interpretability of ML algorithms applying XAI using the dataset called USNW-NB15. Here's what were contributed:

- To improve trust management that is comprehensible to human professionals, the XAI concept was tackled. To do so, feature importance and some machine learning models were used.
- UNSW-NB15 dataset was balanced, and the performance metrics were compared to the imbalanced dataset. Some performance metrics improved because of making the dataset balanced.
- The features extracted from the models using XAI for IDS to enhance human interpretability were interpreted for better explainability.

This paper (4) shows the integration of ML, DL and XAI for intrusion detection systems.. Here they have used ML models such as KNN, SVM, GB, LR, DT and gained the highest accuracy is 99.97% for gradient boosting. They extracted a total of 9 features from 88 features from the CICDDos-2019 dataset. For the explainability of the model they have used SHAP and LIME XAI model. The main challenges they have faced are to increase accuracy and models must be utilized in real-world circumstances. In this paper (5) they worked on the detection of potential cyber-attack using machine learning and XAI technique using SHAP. Two balanced datasets used for different machine learning classification such as DT, random forest, XGBoost, SVM, KNN etc. From the given results XGBoost, DT and random forest's accuracy was so close which is 99.98%, 99.88% and 99.99%. They used SHAP for explainability in XGBoost model which shows the interpretability and understandability of IDS. Has scope to implement LIME and ELI5 with different dataset (6). A framework called FAIXID improves the understandability and explainability of intrusion detection alerts by leveraging XAI and data cleaning techniques. In their paper, they used Premodeling XAI and Post modeling XAI. Framework evaluation evaluated whether the framework could fulfill the goal or not and if data cleaning can increase the explainability of IDS results or not. So, this framework increases the explainability from 0.066 to 0.2056.There is scope for finding feature importance in their work. This research (7) effort suggests a hybrid intrusion detection system where they paid close attention to interpretability and explainability in addition to ml support. To populate the knowledge base with ML-suggested rules and maintain an environment that is understandable and interpretable for the IDS, they have primarily blended two approaches. The machine learning model was implemented Decision tree with Scikit-Learn. This work can expand with using other promising ML models and explained. This paper (8) created self-organizing maps (SOM) based on the X-IDS system, which is able to produce explanatory visualizations. Self-organizing maps map basically works to transform one-dimensional data to higher or two dimensions using this technique in the IDS to get both local and global explanations. Producing reliable IDS and visualization such as feature importance, U-matrices, and feature heatmaps was their primary goal. They have achieved 91% and 80% accuracy in applying datasets on NSL-KDD and CIC-IDS-2017 respectively. It is possible to improve the architecture to get higher accuracy. They have scope to work with all malicious attacks instead of DDoS only. This paper (9) shows experimental analysis to enhance the AI-based In-Vehicle (IV-IDS) IDS as the primary goal. It mainly tackles the problem of interpreting anomaly detection for cybersecurity in automobiles and false alarms where it presents a visualization-based explanation strategy called "VisExp" that, when given to experts, greatly enhanced their level of faith in the system in comparison to explanations based on rules. XAI integration with the IoT is explored in the study (10) where overview is done on the State-of-the-Art and Future Directions. It highlights the necessity of openness in AI choices, especially for IoT applications where also highlights the difficulties and potential prospects in this area. The use of ensemble methods such as DT and RF with SHAP for IDS is the main topic of this paper (11) where it tackles the difficulty of elucidating AI judgments in IDS. They use SHAP to produce outcomes that are comprehensible and easy to grasp. Effective cybersecurity decision-making is made possible by this method, which interprets ensemble tree model results and optimizes them as necessary. They have used three different dataset such as IoTID20, NF-BoT-IoT-v2, NF-ToN-IoT-v2 and obtained 100% accuracy for DT ,100% and 99.69% for E-Graphs AGE They can use LIME, ELI5 with SHAP to get

better comparison. This paper (12) proposed an IDS using ML, where they have implemented classifiers-Bayes Net, J48(Decision tree), RF, Random Tree. KDDCup99 Test datasets were used to determine the accuracy for various types of attacks. They measured precision, recall and f-measure and got 97.4 % precision for Random Tree and 92.3% f-measure. They have scope to use real world dataset to gain more accurate results.

In Section 2, a detailed explanation of the methodology is provided. The proposed methodology and an explanation of the XAI techniques can be observed from there. The result analysis is presented in Section 3. The paper is concluded in Section 4.

2. METHODOLOGY

An outline of the proposed IDS model architecture is shown in Fig. 3. Here, the model's steps are displayed one after the other. The pre-processing steps include data cleaning, normalization, balancing, and transformation which are displayed here for the UNSW-NB15 dataset. Then in the model training, the classifiers Random Forest, SVM, CatBoost are trained using the UNSW-NB15 dataset. Besides upon these 3 supervised classifiers, ensemble learning method is applied. The performances of ensemble approaches using stacking and voting classifiers are also observed. Basically, a performance comparison is obtained from the classification report achieved by training these classifiers. After doing so, ELI5 and LIME, which are the XAI approaches to the interpretability of the model are applied to get a better understanding and interpretation of the model. Feature importance which is defined as scores used to determine the relative extent of individual features in a dataset while constructing a predictive model is obtained. Also, model explanation through it is possible. Model explanation is a collection of procedures and techniques that, as opposed to relying just on blind faith, enable users to comprehend and value the output of a ML algorithm. By using XAI, it is possible to explain the model, which will enhance the reliability and interpretability of the end users. ML models are normally black boxes where interpretation is impossible. But through XAI, it is possible to interpret the reasoning behind algorithms showing certain performances. The dataset collection, data processing steps, classifiers utilized for analysis, and proposed methodology covering XAI techniques ELI5 and LIME are the main subjects of Section 2.

A. Dataset

The UNSW-NB15 computer network security dataset (13) is a product of the Cyber Range Lab at the University of New South Wales Canberra, which was made available in 2015, served as the source material for this study. The dataset contains 2,540,044 examples of both normal and abnormal network behavior, generated by the IXIA PerfectStorm program. The dataset consists of three sets of attributes: fundamental, content, and time. In Fig. 3, the intrusion detection model proposed is illustrated.

B. Pre-Processing

Data pre-processing is an essential process for organizing data that is structured as well as unstructured for any kind of data analysis or data mining. It basically converts raw data into suitable formats for machine learning. Applying feature selection and associated techniques can enhance the quality of features within a dataset and the insights obtained. The two primary categories are numerical and categorical features. Categorical Features are features that describe variables with fixed and limited values and Numerical Features are the continuous features that can take on a range of numerical values are known as numerical features (14).

Machines like to work with ordered and organized data like processed texts, photos, and videos. If these are unprocessed it becomes hard for machines to work with them. So, using data pre-processing techniques these data need to be cleaned and processed for usage.

a) Data cleaning: Data cleaning involves removing erroneous, duplicate, or incomplete information from databases and is an essential stage in data management. It is a component of the pre-processing step, where the objective is to fix inaccurate data, eliminate redundant information, and deal with missing or incomplete information.

b) Normalization: To facilitate faster querying and analysis and improve corporate decision-making, data normalization is a procedure that helps establish a uniform data format throughout a system. To provide a more logical storage strategy, it entails rearranging data sets to eliminate unstructured or unnecessary information. To scale values to a common range so they can be compared to other data sets, data normalization formulas are utilized. The formula modifies the variation of the data set between 0 and 1, where the lowest data point has zero and the highest has a one-valued normalized value. The other data points have decimal values ranging from zero to one. Mathematically, the normalizing equation looks like this (15) :

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

c) Balancing: The UNSW-NB15 dataset was used in the study which was a case of imbalanced dataset at first. Random undersampling is used here to make the UNSW-NB15 dataset balanced in this pre-processing

step of dataset analysis. In Fig. 1, it is seen that the class distribution of label 0 and label 1 are not balanced. Before undersampling the class distribution be as Fig. 1.

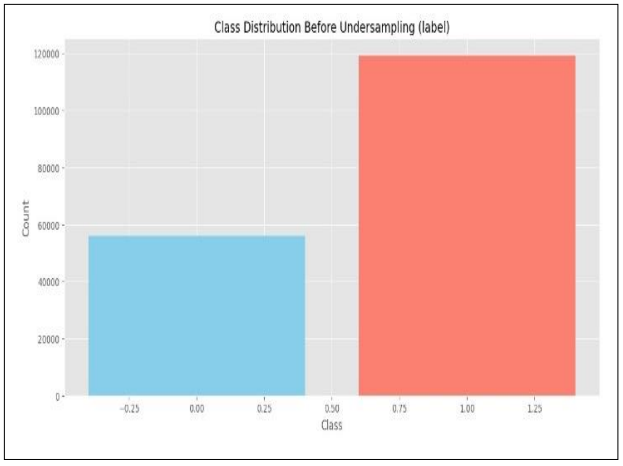

Figure 1. Imbalanced Dataset Class Label

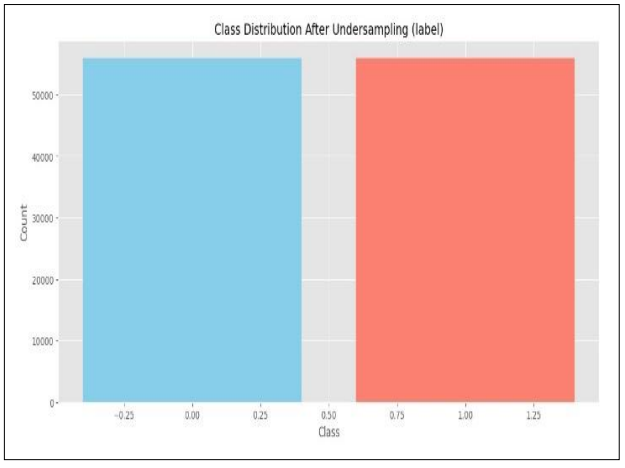

Figure 2. Balanced Dataset Class Label

After doing random undersampling of the majority class, which is label 1, we obtain balanced distribution of class label in Fig. 2.

d) Transformation: The act of transforming, cleaning, and organizing data into a format that can be used for analysis and assist decision-making processes to advance an organization's expansion is known as data transformation. Data munging and data wrangling are other terms for data manipulation which is also the process of transformation of data by modifying it. Data transformation will begin the process of transforming the data into the format needed for analysis and subsequent processes, which we have already begun with data cleansing (16).

C. Model Implementation

To do implementation of the IDS model as Fig. 3, the ML algorithms as follows are to be taken and trained for implementation.

a) Random Forest: Random Forest is an ensemble method to manage supervised categorization. Random forests stem from building decision trees to using training sets and supervised learning techniques to improve the accuracy of the algorithms. The bagging technique is used by Random Forest to construct decision tree ensembles. Random forests generate several decision trees based on random data selections. The primary benefit of random forests is that they make less classification errors (17).

b) *Support Vector Machine:* SVM algorithm aims to find a hyperplane that can discriminate between data points, where N is the number of attributes, to minimize computational time when dealing with millions of samples. This approach lowers the classification risk rather than attempting to achieve the best classification. Utilizing hyperplanes, data points that fall into several groups according to their location on the hyperplane can be categorized. Furthermore, the hyperplane's dimensions are determined by the number of features. A line can only be a hyperplane if it has two input features. The data points that establish a hyperplane's orientation and position are called support vectors. The SVM is the most dependable and effective model and classification technique in the high dimensional feature space for two ad hoc classification issues between two classes (18).

c) CatBoost: A potent machine learning method that has produced exceptional results in a variety of applications is the CatBoost algorithm. However, CatBoost was designed to deal with qualities that are categorical. It can still handle properties that are continuous or numerical. The gradient-boosting decision tree technique now incorporates the cat boost model as a special feature. There are GPU and CPU implementations for CatBoost. It is based on decision trees like other boosting approaches (e.g., XGBoost, LightGBM).The main notion of CatBoost is to sequentially add trees where each new tree tries to fix the faults committed by the previous ones. CatBoost includes various advances, like ordered boosting and an efficient handling of categorical variables, making it stand out in the family of boosting algorithms (19).

d) Ensemble Methods: The ensemble methods that were used in this work are voting and stacking. Voting is one way that combines several techniques to produce improved outcomes. Utilizing this approach, we must first create several categorization models utilizing the dataset for training. In our code the soft voting technique was utilized to achieve the results. Soft voting is used in classification of our research work.

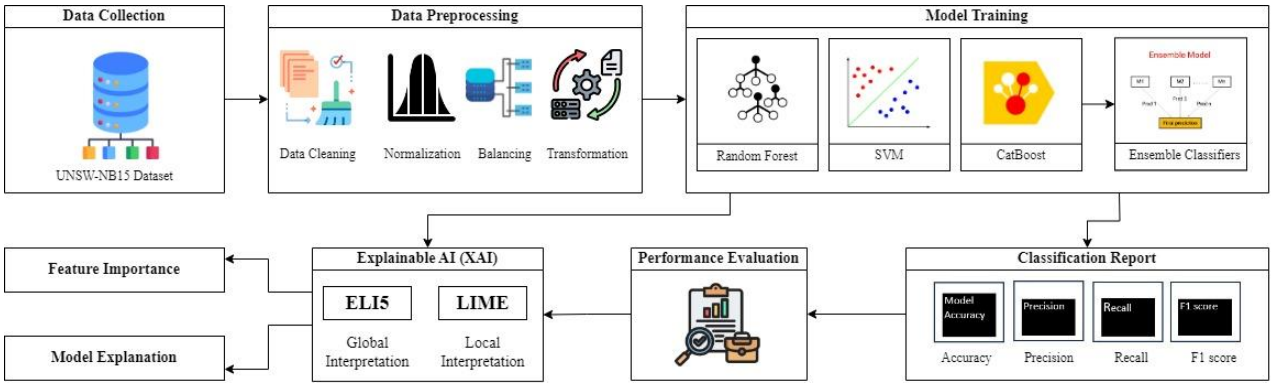

Figure 3. The Intrusion Detection Model Architecture

In this approach, instead of each model in the ensemble voting for a single class (hard vote), they predict the probability of each class. The final output class is determined by averaging the probabilities given by all the models in the ensemble. On the other hand, Stacking is a machine learning ensemble learning technique. The center level classifiers in the stacking ensemble approach are learned using the entire training dataset. The predictions made by the basic learners are sent into the Meta classifier as input and are handled as a fresh dataset (20).

D. Model Explanation using Explainable AI

In our model, we have used two popular techniques in the field of XAI which are ELI5 (Explain Like I'm 5) and LIME (Local Interpretable Model-agnostic Explanations). They contribute to the transparency and interpretability of complicated machine-learning models by shedding light on their decision-making processes (21).

1) LIME: LIME is defined as a framework designed or library that provides explanations for machine learning models' predictions. LIME provides local interpretation, examines each of the model's individual predictions, and tries to explain the model's choices. The way LIME XAI operates is by producing justifications for each forecast.

*2) ELI5:*To enhance the interpretability of ML models, ELI5 robust and intuitive which is a Python library. Individuals can ask questions in the ELI5 subreddit and receive clear, concise answers in return. Regardless of experience level or background, the subreddit aims to make complex subjects understandable to all users. ELI5 provides explanations for each forecast, allowing users to understand the steps of a model.

In the last part of the model of Fig. 3, we see the feature importance and model explanation through which we can do result analysis.

3. RESULT ANALYSIS

The proposed models' behavior and performance were explained using XAI techniques that made use of ELI5 and LIME to facilitate performance improvement. The aim was to build a model that could explain the classification predictions and offer good accuracy. Using the UNSW-NB15 IoT-based network traffic dataset, the performance of the ML classifiers Random Forest, SVM, Catboost, and some ensemble stacking and voting algorithms was assessed through observing their classification reports in subsection A and B. Then, using XAI approaches, the normal and attack prediction probabilities in the ML classifiers were found and explained using ELI5 and LIME in subsection C and D.

A. Classification Report of Imbalanced and Balanced Dataset

The classification report is basically a performance evaluation metric of the machine learning algorithms that evaluates the model by classifying different instances into various classes. It basically evaluates how well the model works by measuring the trained models- Precision, F1 Score, Accuracy and Support. The confusion matrix, a tabular representation of the actual and anticipated results of any machine learning model, is the source of the classification report. A detailed study on the models predictions can be observed by analyzing their classification reports. The classification reports containing the necessary information are illustrated here.

Classification Report for Random Forest Classifier:

	precision	recall	f1-score	support
0	0.78	0.96	0.86	56000
1	0.98	0.87	0.92	119341
accuracy			0.90	175341
macro avg	0.88	0.92	0.89	175341
weighted avg	0.92	0.90	0.91	175341

(a) Imbalanced Dataset

Classification Report for Random Forest Classifier:

	precision	recall	f1-score	support
0	0.88	0.96	0.92	56000
1	0.96	0.87	0.91	56000
accuracy			0.92	112000
macro avg	0.92	0.92	0.92	112000
weighted avg	0.92	0.92	0.92	112000

(b) Balanced Dataset

Figure 4. Classification Report of Random Forest

Classification Report:

	precision	recall	f1-score	support
0	0.72	0.93	0.81	56000
1	0.96	0.83	0.89	119341
accuracy			0.86	175341
macro avg	0.84	0.88	0.85	175341
weighted avg	0.89	0.86	0.87	175341

(a) Imbalanced Dataset

Classification Report for SVM Classifier:

	precision	recall	f1-score	support
0	0.85	0.93	0.89	56000
1	0.92	0.83	0.88	56000
accuracy			0.88	112000
macro avg	0.89	0.88	0.88	112000
weighted avg	0.89	0.88	0.88	112000

(b) Balanced Dataset

Figure 5. Classification Report of SVM

Classification Report:

	precision	recall	f1-score	support
0	0.77	0.98	0.86	56000
1	0.99	0.86	0.92	119341
accuracy			0.90	175341
macro avg	0.88	0.92	0.89	175341
weighted avg	0.92	0.90	0.90	175341

(a) Imbalanced Dataset

Classification Report for CatBoost Classifier:

	precision	recall	f1-score	support
0	0.87	0.98	0.92	56000
1	0.97	0.86	0.91	56000
accuracy			0.92	112000
macro avg	0.92	0.92	0.92	112000
weighted avg	0.92	0.92	0.92	112000

(b) Balanced Dataset

Figure 6. Classification Report of CatBoost

Classification Report:

	precision	recall	f1-score	support
0	0.76	0.97	0.85	56000
1	0.98	0.86	0.91	119341
accuracy			0.89	175341
macro avg	0.87	0.91	0.88	175341
weighted avg	0.91	0.89	0.89	175341

(a) Imbalanced Dataset

Classification Report:

	precision	recall	f1-score	support
0	0.87	0.97	0.91	56000
1	0.96	0.85	0.90	56000
accuracy			0.91	112000
macro avg	0.91	0.91	0.91	112000
weighted avg	0.91	0.91	0.91	112000

(b) Balanced Dataset

Figure 7. Classification Report of Ensemble Stacking classifier

Classification Report:

	precision	recall	f1-score	support
0	0.77	0.98	0.86	56000
1	0.99	0.86	0.92	119341
accuracy			0.90	175341
macro avg	0.88	0.92	0.89	175341
weighted avg	0.92	0.90	0.90	175341

(a) Imbalanced Dataset

Classification Report:

	precision	recall	f1-score	support
0	0.87	0.98	0.92	56000
1	0.98	0.86	0.91	56000
accuracy			0.92	112000
macro avg	0.93	0.92	0.92	112000
weighted avg	0.93	0.92	0.92	112000

(b) Balanced Dataset

Figure 8. Classification Report of Ensemble Voting classifier

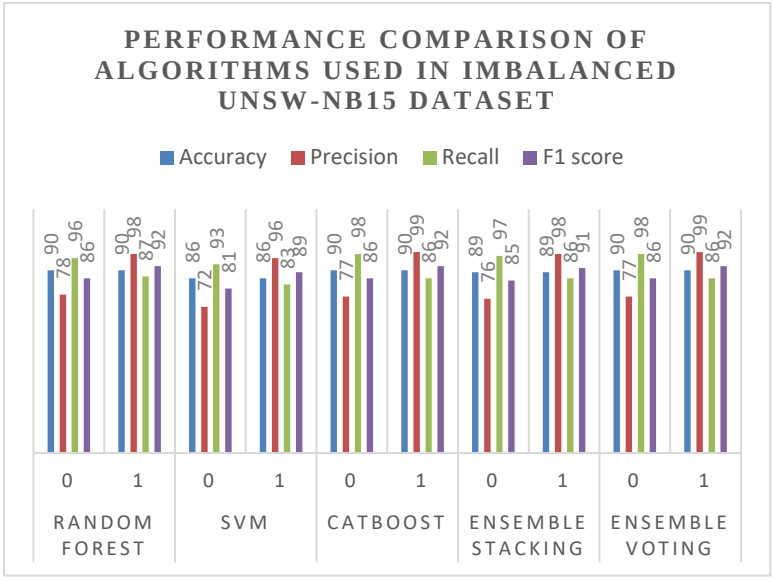

Figure 9. Classification Report of Voting for Balanced Dataset

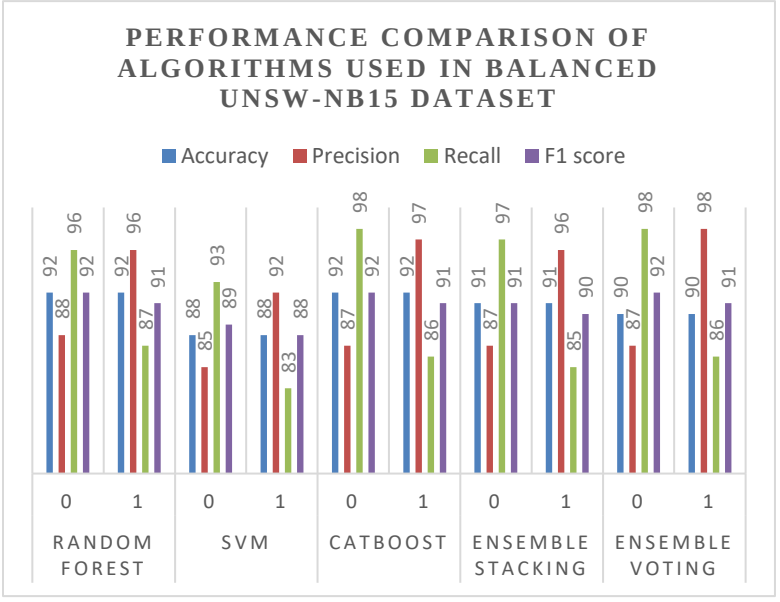

Figure 10. Classification Report of Voting for Balanced Dataset

B. Performance Comparison

The performance comparison of all these algorithms in case of these algorithms being trained on the imbalanced and balanced UNSW-NB15 dataset is represented in Fig. 9 and Fig. 10. The performance evaluation measures - Recall, Accuracy, Precision, and F1 score for the classifiers that were used—RF, SVM, CatBoost, Ensemble Stacking and Ensemble Voting (Soft Voting) are presented in Fig. 9 and Fig. 10. These are shown for both the cases of imbalanced and balanced dataset. In the case of Fig. 10, i.e. for balanced UNSW-NB15 dataset , 90 percent is the highest accuracy that comes amongst all classifiers. All these cases give accuracy of 90 or somewhat near that as observed in Figure 9 and Fig. 10. Similarly, the maximum accuracy that the classifiers—RF and CatBoost—can provide for the imbalanced UNSW-NB15 dataset is 92 percent in Fig. 9. Compared to Stacking (91 percent), Voting (90 percent), and other algorithms of classification, the ensemble method classifiers yield less accuracies. Besides, we see improvement in the precision values of class 0 and class 1 from imbalanced to balanced datasets. Accuracies and other metrics also improve in balanced dataset. When predicting different characteristics of the dataset, these performance metrics provide insightful information about how well-performing and reliable each algorithm is.

C. XAI ELI5 Implementation for Imbalanced and Balanced Dataset

The top features in the dataset, Permutation Importance (PI) or Mean Decrease Accuracy (MDA), which when run through any classifier yields some accuracy, were identified using the ELI5 PI toolkit and the Scikit Learn library. To improve the explainability of classification prediction, it is necessary to find the permutation importance module, which computes feature importance by observing how score drops when a feature is absent. And such is done for the Random Forest applied in the imbalanced and balanced UNSW-NB15 dataset to observe the weight of each feature.

From Fig. 11(a) and Fig. 11(b) it can be observed that the features ‘‘ ct_dst_src_ltm’’, ‘‘ ct_state_ttl’’ and “sttl” are the top features. Amongst the top 3 features ,we find that two of the features matches. In the case of the other classifiers those are non-tree based. ELI5 is not applied to those other algorithms, rather we use LIME to find the interpretations. In the case of ML algorithms, it is one of the essential Python libraries that will aid in the model's interpretability and explainability. It helps in the overall decision-making process of any algorithm.

Weight	Feature
0.0220 ± 0.0005	sttl
0.0171 ± 0.0005	ct_dst_src_ltm
0.0063 ± 0.0002	dttl
0.0052 ± 0.0004	ct_srv_dst
0.0026 ± 0.0001	ct_srv_src
0.0018 ± 0.0002	ct_state_ttl
0.0009 ± 0.0002	swin
0.0005 ± 0.0001	dmean
0.0004 ± 0.0001	spkts
0.0002 ± 0.0001	ct_src_ltm
0.0002 ± 0.0001	djit
0.0002 ± 0.0001	ct_dst_ltm
0.0000 ± 0.0000	is_sm_ips_ports
0.0000 ± 0.0001	ct_src_dport_ltm
0.0000 ± 0.0000	ct_flw_http_mthd
0 ± 0.0000	is_ftp_login
0 ± 0.0000	ct_ftp_cmd
0 ± 0.0000	trans_depth
-0.0000 ± 0.0000	response_body_len
-0.0001 ± 0.0001	sloss
… 19 more …	

(a) Imbalanced Dataset

Weight	Feature
0.0220 ± 0.0004	sttl
0.0206 ± 0.0009	ct_dst_src_ltm
0.0048 ± 0.0003	ct_srv_dst
0.0041 ± 0.0002	dttl
0.0037 ± 0.0002	smean
0.0036 ± 0.0003	sload
0.0023 ± 0.0002	ct_srv_src
0.0020 ± 0.0005	dbytes
0.0012 ± 0.0001	ct_state_ttl
0.0011 ± 0.0004	sbytes
0.0004 ± 0.0001	ct_dst_ltm
0.0004 ± 0.0003	ct_dst_sport_ltm
0.0004 ± 0.0000	spkts
0.0003 ± 0.0000	ct_src_ltm
0.0002 ± 0.0001	ct_src_dport_ltm
0.0001 ± 0.0000	djit
0.0001 ± 0.0002	dmean
0.0001 ± 0.0002	swin
0.0000 ± 0.0000	is_sm_ips_ports
0 ± 0.0000	is_ftp_login
… 19 more …	

(b) Balanced Dataset

Figure 11. ELI5 Permutation Importance for Random Forest

D. XAI LIME Implementation for Imbalanced and Balanced Dataset

LIME is applied in the algorithms and the visual dashboard indicates which features and their weights brought the classification overall to be considered ‘Normal’ or ‘Attack’ in case of the imbalanced and balanced UNSW-NB15 network traffic dataset. This is useful to find various prediction probabilities and classification tasks of different ML algorithms.

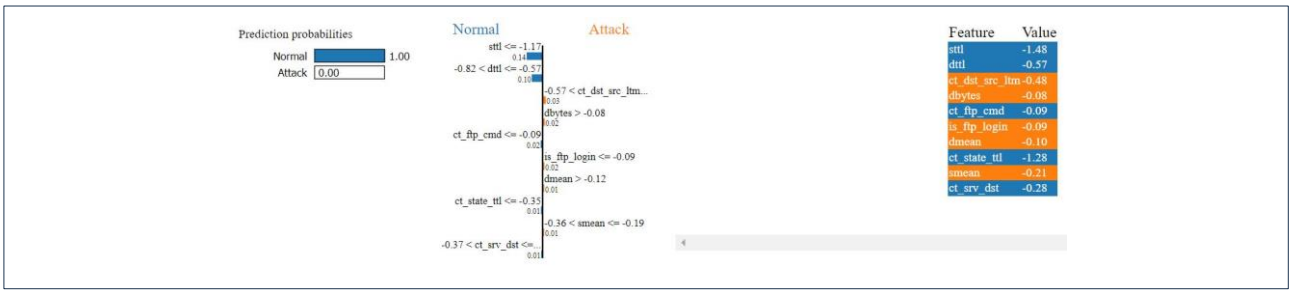

Figure 12. LIME Explanation of instance i =1653 of Random Forest in Imbalanced UNSW-NB15 Dataset

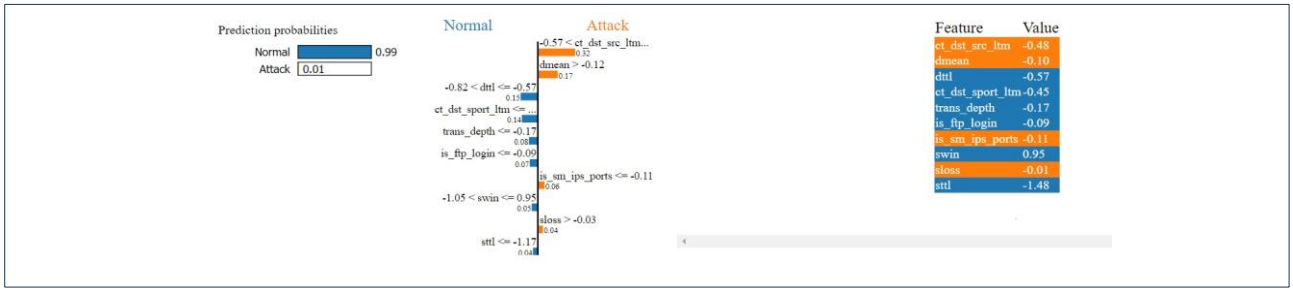

Figure 13. LIME Explanation of instance i =1653 of SVM in Imbalanced UNSW-NB15 Dataset

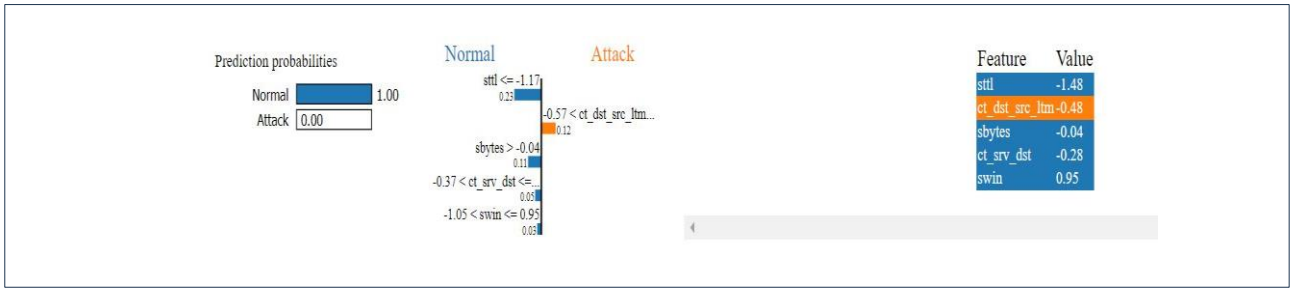

Figure 14. LIME Explanation of instance i =1653 of CatBoost in Imbalanced UNSW-NB15 Dataset

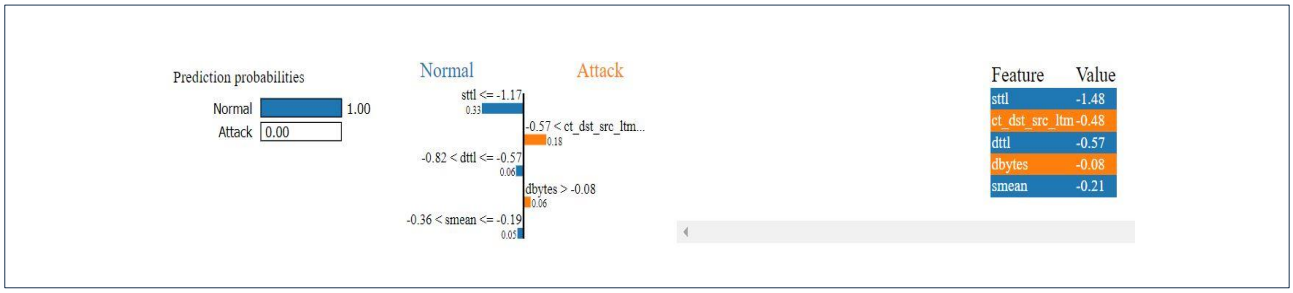

Figure 15. LIME Explanation of instance i =1653 of Stacking classifier in Imbalanced UNSW-NB15 Dataset

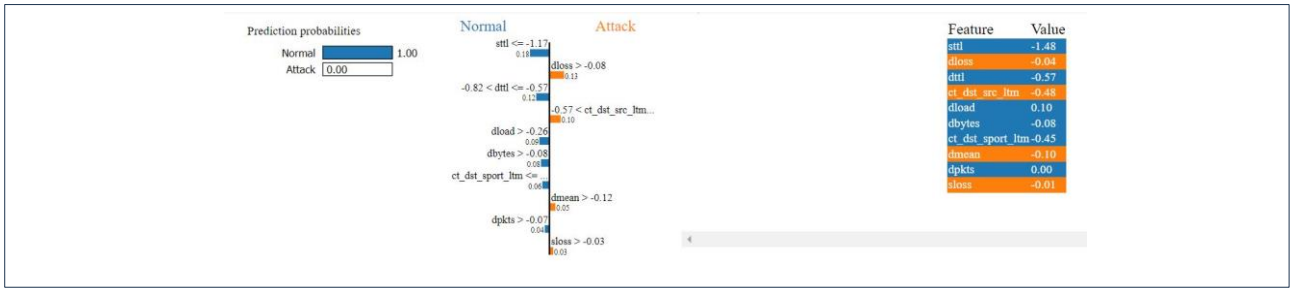

Figure 16. LIME Explanation of instance i =1653 of Voting classifier in Imbalanced UNSW-NB15 Dataset

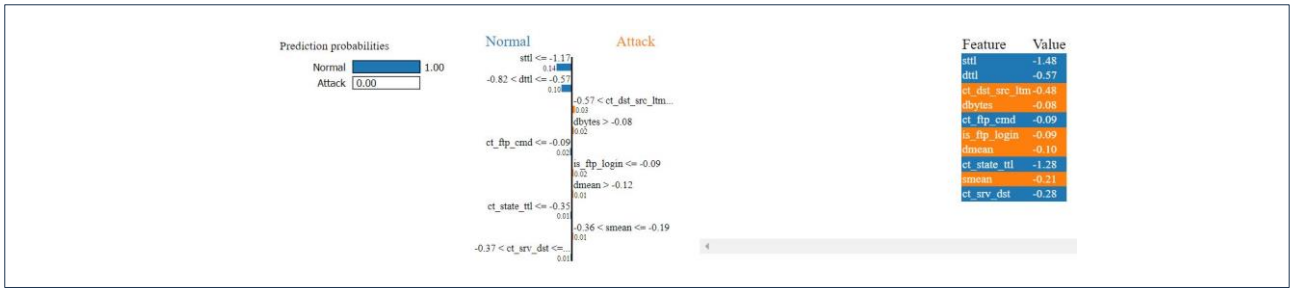

Figure 17. LIME Explanation of instance i =1653 of Random Forest in Balanced UNSW-NB15 Dataset

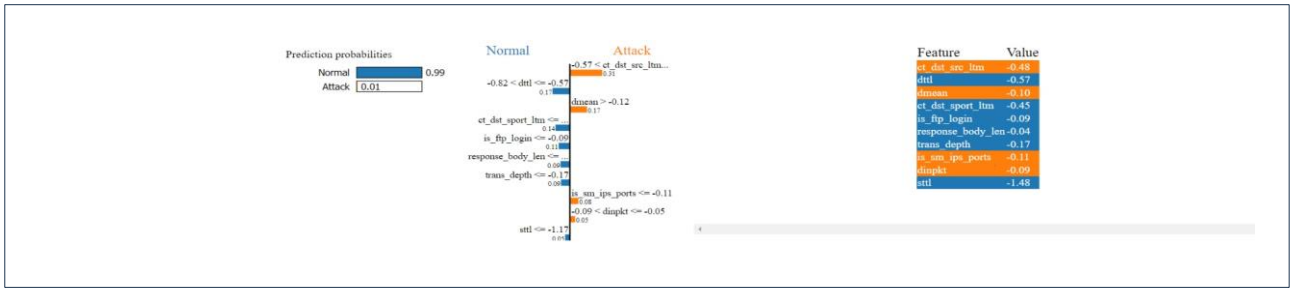

Figure 18. LIME Explanation of instance i =1653 of SVM in Balanced UNSW-NB15 Dataset

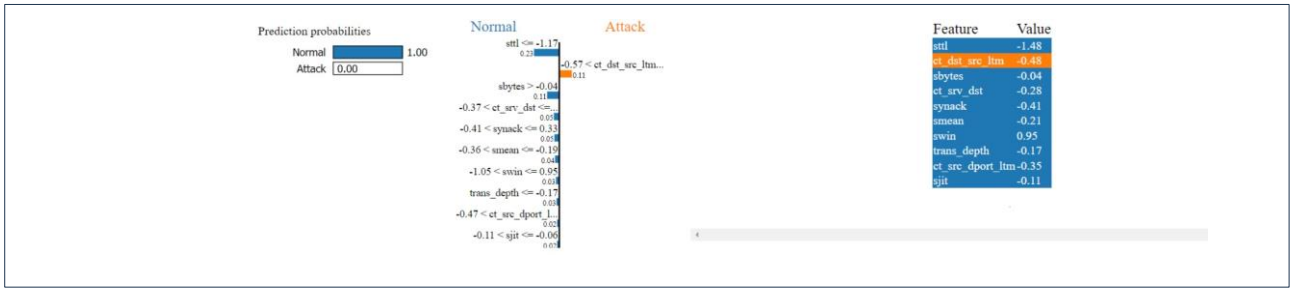

Figure 19. LIME Explanation of instance i =1653 of CatBoost in Balanced UNSW-NB15 Dataset

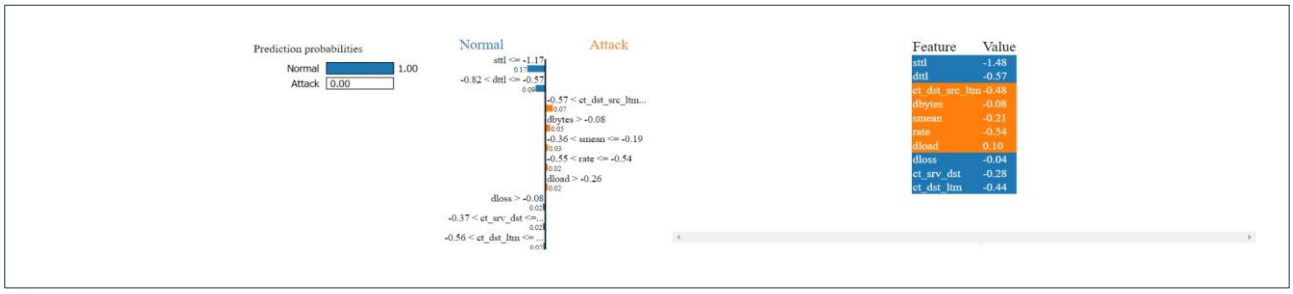

Figure 20. LIME Explanation of instance i =1653 of Stacking in Balanced UNSW-NB15 Dataset

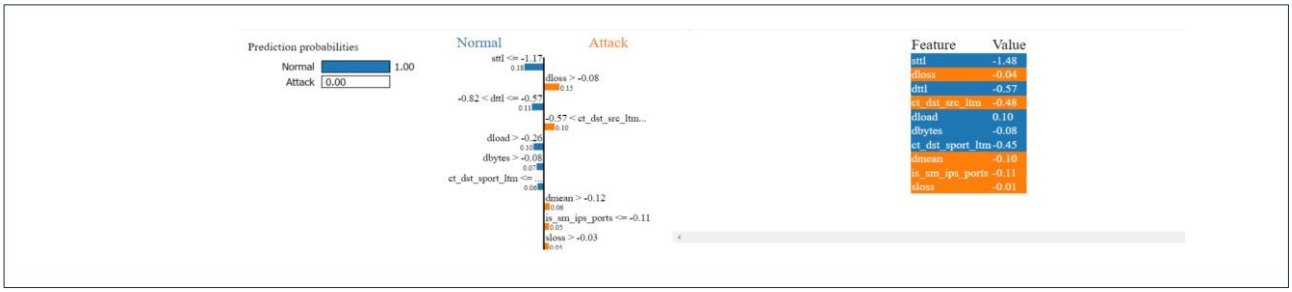

Figure 21. LIME Explanation of instance i =1653 of Voting classifier in Balanced UNSW-NB15 Dataset

The visual dashboards obtained from LIME reveal which feature impact what category of prediction, and for which weight that prediction is obtained. The predictions may be 'Normal' or 'Attack' type. The features influencing 'Normal' prediction are denoted by blue and the features that influence the 'Attack' prediction are denoted by orange in the visual dashboard of it. The features and the values for which the prediction probabilities can be determined are also present in the visual dashboard of LIME which makes it easier to interpret and comprehend.

4. CONCLUSION

To enhance explainability in the establishment of IoT network security through various IDS, this work uses XAI to leverage ML algorithms, also known as "black boxes," where the reasoning or logic behind the output predictions is not always clear. Because of their innate domain knowledge, human analysts continue to be essential for resource allocation and cybersecurity strategy development, even in the face of increasingly complex machine learning models used for IoT network traffic security through IDSs. In this work, the UNSW-NB15 dataset is used as the basis for training a range of classifiers, including Random Forest, SVM and CatBoost. In addition, voting and stacking ensemble classifiers are employed for additional analysis. Subsequently, we employ XAI on those trained classifiers to explain the decisions displayed by those classifiers be viewed in a more comprehensible and detail-oriented manner, thereby facilitating human comprehension of the outcomes or decisions. XAI is employed and implemented using the python libraries ELI5 and LIME. This is how XAI was applied to the ML algorithms to assess its performance and this proceeds to demystify the IoT network intrusion detection. The proposed systems limitation may include the fact that the algorithms took a considerable amount of time to get trained and yield the results. Another limitation for this research work might be the models not keeping up with the ever-evolving dynamic nature of network intrusion attacks. The amount of new security breach attacks could overload the model, making it difficult for it to keep up with them. This paper's primary goal is to use XAI to assess the model's performance. Therefore, it is limited that the emphasis is on XAI rather than the necessity to simplify the model's performance.

References

1. Bace R, Mell P. NIST Special Publication on Intrusion Detection Systems Intrusion Detection Systems.

2. Tiwari R. Explainable AI (XAI) and its Applications in Building Trust and Understanding in AI Decision Making.

INTERANTIONAL J Sci Res Eng Manag. 2023 Jan 27;07(01).

3. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI-Explainable artificial intelligence. Sci Robot. 2019;4(37).

4. Siganos M, Radoglou-Grammatikis P, Kotsiuba I, Markakis E, Moscholios I, Goudos S, et al. Explainable AI-based Intrusion Detection in the Internet of Things. ACM Int Conf Proceeding Ser. 2023;

5. Mallampati SB, Seetha H. A Review on Recent Approaches of Machine Learning, Deep Learning, and Explainable Artificial Intelligence in Intrusion Detection Systems. Vol. 17, Majlesi Journal of Electrical Engineering. Islamic Azad University; 2023. p. 29–54.

6. Liu H, Zhong C, Alnusair A, Islam SR. FAIXID: A Framework for Enhancing AI Explainability of Intrusion Detection Results Using Data Cleaning Techniques. J Netw Syst Manag [Internet]. 2021;29(4):1–30. Available from: https://doi.org/10.1007/s10922-021-09606-8

7. Dias T. A Hybrid Approach for an Interpretable and Explainable Intrusion Detection System.

8. Shtayat MM, Hasan MK, Sulaiman R, Islam S, Rehman AU. An Explainable Ensemble Deep Learning Approach for Intrusion Detection in Industrial Internet of Things. IEEE Access. 2023;11(October):115047–61.

9. Lundberg H, Mowla NI, Abedin SF, Thar K, Mahmood A, Gidlund M, et al. Experimental Analysis of Trustworthy In-Vehicle Intrusion Detection System Using eXplainable Artificial Intelligence (XAI). IEEE Access. 2022;10(October):102831–41.

10. Jagatheesaperumal SK, Pham QV, Ruby R, Yang Z, Xu C, Zhang Z. Explainable AI Over the Internet of Things (IoT): Overview, State-of-the-Art and Future Directions. IEEE Open J Commun Soc. 2022;3(October):2106–36.

11. Le TTH, Kim H, Kang H, Kim H. Classification and Explanation for Intrusion Detection System Based on Ensemble Trees and SHAP Method. Sensors. 2022;22(3):1–28.

12. Maheswaran N, Bose S, Logeswari G, Anitha T. Hybrid Intrusion Detection System Using Machine Learning Algorithm. Lect Notes Networks Syst. 2023;572(1):333–46.

13. Moustafa N, Slay J. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: 2015 Military Communications and Information Systems Conference, MilCIS 2015 - Proceedings. Institute of Electrical and Electronics Engineers Inc.; 2015.

14. Suad A. Alasadi, Wesam S. Bhaya. Review_of_Data_Preprocessing_Techniques. Vol. 12, Jurnal of Engineering and Applied Sciences. 2017. p. 4102–7.

15. Mohammed Shantal, Zalinda Othman AAB. SS symmetry A Novel Approach for Data Feature Weighting Using. 2023;

16. Fan C, Chen M, Wang X, Wang J, Huang B. A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. Front Energy Res. 2021;9(March):1–17.

17. Wali S, Khan IA, Member S. Explainable AI and Random Forest Based Reliable Intrusion Detection system. techarXiv [Internet]. 2021; Available from: https://www.techrxiv.org/articles/preprint/Explainable_AI_and_Random_Forest_Based_Reliable_Intrusion_Detection_system/17169080

18. Sadqui A, Ertel M, Sadiki H, Amali S. Evaluating Machine Learning Models for Predicting Graduation Timelines in Moroccan Universities. Int J Adv Comput Sci Appl. 2023;14(7):304–10.

19. Ghori KM, Abbasi RA, Awais M, Imran M, Ullah A, Szathmary L. Performance Analysis of Different Types of Machine Learning Classifiers for Non-Technical Loss Detection. IEEE Access. 2020;8:16033–48.

20. Raihan-Al-Masud M, Mustafa HA. Network Intrusion Detection System Using Voting Ensemble Machine Learning. 3rd IEEE Int Conf Telecommun Photonics, ICTP 2019. 2019;(December 2019).

21. Sivamohan S, Sridhar SS. An optimized model for network intrusion detection systems in industry 4.0 using XAI based Bi-LSTM framework. Neural Comput Appl. 2023 May 1;35(15):11459–75.

Tasfia Zaima received her B.Sc. in Information and Communication Engineering degree from Bangladesh University of Professionals. She is currently pursuing a master's degree in the same subject and University. Her interest in research lies in the field of cybersecurity, AI, and natural language processing.

Tabassum Ibnat Ena is a graduate of Information and Communication Engineering from Bangladesh University of Professionals in Bangladesh. Her current research interest is in AI and Cybersecurity.

Md. Tamim Ikbal was born in Barisal, Bangladesh. He received B.Sc. in Information and Communication Engineering from Bangladesh University of Professionals. His research interest lies in the field of cybersecurity and source code vulnerability analysis.

Abu Sayed Md. Mostafizur Rahaman received his PhD degree from the Department of Computer Science and Engineering of Jahangirnagar University in Savar, Dhaka, Bangladesh in 2014. He obtained his B.Sc. degree in Electronics and Computer Science from Jahangirnagar University, Savar, Dhaka, Bangladesh in 2003, and his M.Sc. degree in Information Technology (INFOTECH) in Embedded System Engineering from Stuttgart University in Stuttgart, Germany in 2009.Since 2004, he is a faculty member having current Designation "Professor" in the Department of Computer Science and Engineering of Jahangirnagar University, Savar, Dhaka, Bangladesh. He completed an industrial internship as a Trainee Engineer at BOSCH, the largest automobile company in Germany, while pursuing a graduate degree in embedded systems. His current research focuses on Digital Forensics, Cryptography, IoT, web Security and S/W Systems.